DOCUMENT RESUME

ED 147 386 UD 017 541

AUTHOR	David, Jane L.; Pelavin, Sol H.
TITLE	Research on the Effectiveness of Compensatory Education Programs: A Reanalysis of Data.
INSTITUTION	Stanford Research Inst., Menlo Park, Calif. Educational Policy Research Center.
SPONS AGENCY	National Inst. of Education (DHEW), Washington, D.C.
BUREAU NO	URU-4425
PUB DATE	Sep 77
CONTRACT	400-75-0066
NOTE	77p.
EDRS PRICE	MF-$0.83 HC-$4.67 Plus Postage.
DESCRIPTORS	Compensatory Education; *Compensatory Education Programs; *Data Analysis; Definitions; Measurement Goals; *Measurement Techniques; *Program Effectiveness; *Program Evaluation; *Time Factors (Learning)
IDENTIFIERS	*Elementary Secondary Education Act Title I

ABSTRACT

This study attempts to clarify some of the issues involved in defining the effectiveness of compensatory education programs. It discusses the extent to which conclusions about the effectiveness of compensatory education programs are affected by two major components of an evaluation: the period of time on which the evaluation is based and the standard against which the program's effectiveness is judged. It is argued that the philosophy of compensatory education suggests that evaluations should measure program effectiveness over a period of time longer than the school year; in other words, that evaluations should assess the extent to which effects are sustained. Achievement gains for several programs, based on at least two periods of time, are calculated. The primary finding of these analyses is that conclusions about program effectiveness, regardless of what standard is used, are greatly influenced by the period of time over which the program is judged. Specifically, it is shown that the inclusion of the summer months in the evaluation can substantially reduce estimates of achievement and often reverse positive judgements of program effectiveness. The data demonstrate that programs can show evidence of sustained effects. (Author/AM)

UD

RESEARCH ON THE EFFECTIVENESS OF COMPENSATORY EDUCATION PROGRAMS: A REANALYSIS OF DATA

Final Report

September 1977

SRI Project URU-4425

CONTRACT 400-75-0066

Prepared for:

National Institute of Education
Department of Health, Education and Welfare
Washington, D.C. 20208

EDUCATIONAL POLICY RESEARCH CENTER

333 Ravenswood Ave.
Menlo Park, California 94025
(415) 326-6200
Cable: STANRES, Menlo Park
TWX: 910-373-1246

SRI International

SRI International

EDUCATIONAL POLICY RESEARCH CENTER

RESEARCH ON THE EFFECTIVENESS OF COMPENSATORY EDUCATION PROGRAMS: A REANALYSIS OF DATA

Final Report

September 1977

SRI Project URU-4425

Project Director: SOL H. PELAVIN

Authors: JANE L. DAVID
SOL H. PELAVIN

Prepared for:

National Institute of Education
Department of Health, Education and Welfare
Washington, D.C. 20208

The conclusions and recommendation in this report are those of the Contractor and do not necessarily reflect the views of the U.S. Department of Health, Education and Welfare or any other agency of the government.

333 Ravenswood Ave.
Menlo Park, California 94025
(415) 326-6200
Cable: STANRES, Menlo Park
TWX: 910-373-1246

ACKNOWLEDGEMENTS

We would first like to express our indebtedness to Susan Bryson on whom we relied heavily throughout the project. She assumed the major responsibility for preparing, analyzing and summarizing the data—a task of huge proportions. She also assisted in numerous other ways from obtaining additional information from districts to proofreading the reports. We are deeply grateful to the staff of the districts whose data we obtained. Without their willingness to cooperate and donate valuable time as well as providing an important local perspective, the study would have been impossible. Similarly we owe gratitude to the project monitor, Dr. Joy Frechtling of the National Institute of Education, whose patience and constructive criticism facilitated the conduct of the study and greatly improved the final report. Finally, we thank Arlowayne Brenno and Bonnie Miller whose skill, kindness and good humor were essential in the creation of this report.

CONTENTS

ILLUSTRATIONS

TABLES

I INTRODUCTION

Evaluations of compensatory education programs have raised more questions than they have answered. Much of the resulting confusion is inevitable in a field that is young and expanding. The confusion is exacerbated, however, by a lack of adequate attention to how program effectiveness is defined, particularly in the context of the philosophy of compensatory education. This study is an attempt to clarify some of the issues involved in defining the effectiveness of compensatory education programs.

The work reported here concerns the extent to which conclusions about the effectiveness of compensatory education programs are affected by two major components of an evaluation: the period of time on which the evaluation is based and the standard against which the program's effectiveness is judged. We argue in particular that the philosophy of compensatory education suggests that evaluations should measure program effectiveness over a period of time longer than the school-year; in other words, that evaluations should assess the extent to which effects are sustained. Therefore, we calculate achievement gains for several programs based on at least two periods of time: the traditional fall pretest to spring posttest (school-year) evaluation period and a 12-month, fall-to-fall period that includes the summer following the program.

We then draw conclusions about program effectiveness based on three standards for success and compare the conclusions for the different time periods. These standards are derived from those previously used in evaluations of compensatory education programs and use the norms of standardized tests as the frame of reference. Two of the standards are expressed in the metric of grade equivalents: a rate of gain of one grade-equivalent month for each month in the program and an annual rate of 8 months. The third standard is a gain of ten percentile points. In the absence of information on the expected achievement of disadvantaged

students without compensatory education experience, we do not select a "best" standard, but rather demonstrate the extent to which conclusions about effectiveness differ according to the standard and the period of time used.

Section II provides an extended discussion of the rationale for measuring sustained effectiveness and for our choice of standards. In Section III, we present a description of the search for appropriate data and each data set obtained. Section IV contains the results of the primary analyses. Section V presents supplemental analyses, and our conclusions are presented and discussed in Section VI.

To simplify the presentation in the text, we have relegated a large number of tables and detail to appendices which are bound separately. This material is referenced throughout the text.

II RATIONALE

Research on compensatory education programs has failed to produce a widely accepted definition of program effectiveness. In fact, research and evaluation are rarely conducted with a clear definition of "success." Researchers and practitioners define effectiveness in a number of ways, ranging from vague statements such as "better than expected" to more sophisticated statements of a required magnitude of change. pose of this work is not to develop a single definition of effectiveness, but to demonstrate how different definitions of effectiveness can lead to different conclusions about program success or failure. Instead of developing a specific definition with limited application, we specify the major ingredients necessary for a definition. In particular, we concentrate on two major components of a definition of effectiveness: the period of time on which the evaluation is based and the standard against which the program is judged.

We begin with the assertion that effectiveness should be defined in the context of the goals of compensatory education programs, and that these goals should determine what should be measured and when the measurements should occur. We have chosen to restrict ourselves to one frequently stated and often measured goal of compensatory education, the improvement of cognitive skills as measured by standardized achievement tests. In the remainder of this section, we discuss the period of time on which an evaluation is based and the standards that we employ in judging program effectiveness.

Period of Time

A fundamental assumption of compensatory education is that greater achievement can change the academic future of disadvantaged students, in turn enhancing their "life chances." Therefore, one of the goals of compensatory education is to increase the achievement of disadvantaged

students. In order to improve students' futures, this increase in achievement should be evident subsequent to participation in a compensatory education program. At a minimum, an increase in achievement should persist over the summer following a school-year program. However, evaluations of compensatory education in general, and of Title I of the Elementary and Secondary Education Act (ESEA) in particular, have not included estimates of sustained achievement. Instead, judgments of program success have been based on students' achievement during the school year; that is, on a spring posttest score adjusted in some way for the preceding fall pretest score.

We hypothesized that evaluations based on measures of sustained achievement would lead to different conclusions than evaluations based on school-year (fall-to-spring) achievement. Specifically, we hypothesized that evaluations based on a fall-to-fall period, by virtue of including the summer months, would result in smaller achievement gains than traditional school-year evaluations. We were led to this hypothesis in part by studies that compare the achievement rates of disadvantaged students during the school year and during the summer (Hayes and Grether, 1969; Heyns, 1976; Murnane, 1975). These studies, while extremely limited, present some evidence that disadvantaged students achieve at a slower rate than expected over the summer. Both conventional wisdom and the standardization procedures of achievement tests assume that the rate of achievement for all students is slower during the summer than during the school year. The grade-equivalent scale defines the rate of achievement of the 50th percentile student as 9 months over the 9-month school year and 1 month over the 3-month summer. Hence the summer rate is assumed to be one-third the school-year rate. This pattern of achievement is presumed to be the same for both advantaged and disadvantaged students: all students are assumed to gain over the summer but at a slower rate than over the school year. The studies cited above suggest that this is not the case for disadvantaged students. In fact, disadvantaged students may have no gain over the summer or may even lose.

The development of the hypothesis was also influenced by the fact that evidence of success of Title I students during the school year was

not supported by other sources of data. Specifically, State Title I evaluations show that students in Title I programs achieve at much higher rates than expected* during the school year. This finding is not supported, however, by data from statewide testing programs. Since the advent of Title I, there are no detectable increases in the scores of those most likely to be Title I participants--the low-percentile students (Thomas and Pelavin, 1976).

Together, these findings suggest that large achievement gains produced by compensatory-education programs over the school year may be followed by corresponding achievement losses over the summer. If such summer losses occur, whether or not they are proportional to school-year gains, evaluations including the summer months will result in smaller achievement gains than evaluations based on the traditional fall-to-spring time period.

Consequently, one major goal of our study was to compare achievement gains for several programs based on different periods of time. However, the period of time used in an evaluation is not the only component that determines whether or not a program is effective. There must also be a standard against which achievement gains are judged. Therefore, the second goal of our work was to illustrate the extent to which conclusions about program effectiveness are affected by the standard used. The standards that we applied and the rationale for using them are described below.

Three Standards for Success

A major problem in the evaluation of compensatory education programs is the lack of information on the expected achievement of disadvantaged students not participating in compensatory education. To determine what portion of an achievement gain is directly attributable to a compensatory

*The expectation is based on the unofficial Title I standard for success, which is one grade-equivalent month gained for each month in the program.

education program, the evaluator must have some notion of what would have happened to students' achievement had they not participated in the program. There is not a large body of data on educationally disadvantaged students who have not been in compensatory education programs. And as more educationally disadvantaged students participate in compensatory programs, such "baseline" data become more difficult to obtain. In the absence of such baseline data, evaluators are faced with a choice among several less than satisfactory alternatives such as using various types of "control" groups or using the norms of standardized tests as the frame of reference.

In evaluations of Title I programs, the use of standards derived from standardized test norms is by far the most common approach. This is partly because the standards, which are expressed in grade-equivalents or percentiles, can be applied across different tests and thus used in aggregating data for national purposes. One such standard that has been applied in the past by the U.S. Office of Education (USOE) is an average achievement rate of one grade-equivalent month per month during the school year. A second is a variation of the standard suggested in one of the recently adopted USOE evaluation models: a percentile increase equivalent to one-third of the standard deviation of the norm group. A third, in the language of grade-equivalents, is in fact empirically based: an achievement gain of 8 grade-equivalent months. The genesis and characteristics of each of these standards are discussed below.

Month-for-Month Standard

Procedures for developing the grade-equivalent scale vary somewhat from one test publisher to another, but all tests define the achievement rate for the average or 50th percentile student to be one grade-equivalent month per month during the 9-month school year and one grade-equivalent month over the 3-month summer. The month-for-month standard stems from this achievement rate, and its application to compensatory education programs rests on the assumption that a disadvantaged student achieving at the rate of the 50th percentile student is doing better than expected. To demonstrate that this assumption is at least open to question, we

describe in oversimplified fashion the derivation of the grade-equivalent scale.

A standardized achievement test is not one test but a battery consisting of several test levels, each spanning one or more grades. The norming of the battery consists of administering adjacent levels of the test battery in each grade to a sample considered to be nationally representative.* From these raw scores, a scale is developed, spanning all test levels, that allows translation of each raw score into a single metric. The median score at each grade G is assigned the grade-equivalent score of G.X where X is the number of the month of the school year in which the test was standardized. For example, if the test were administered in October (one month into the school year), the median score for third graders would be assigned a grade-equivalent score of 3.1. By assigning the appropriate grade-equivalent score to the median score at each grade, a set of grade-equivalent scores (1.1, 2.1, 3.1, etc.) can be plotted against the scale scores that span all levels of the test. In essence, the omitted grade-equivalents (3.2, 3.3, 3.4, etc.) are interpolated by dividing the distance between consecutive median scores into tenths. Thus the score that is one-tenth of the distance from the third grade median score of 3.1 to the fourth grade median of 4.1 is assigned the value of 3.2, and so on.

Both the development of a scale that spans test levels and the interpolations between median scores entail quite complex mathematical manipulations from the application of Thurstone scaling techniques to the fitting of high-order polynomials. The above description is intended only to provide a sketch of the development of the grade-equivalent scale with the understanding that the actual procedures are quite complicated and vary from test to test.

The point of describing this procedure is to provide an understanding of the empirical basis for the month-for-month standard. In essence, the pattern of achievement described by the grade-equivalent scale is based

---

*This discussion describes standardization procedures based on one test administration.

only on median scores, a grade-equivalent year apart. All other grade-equivalents are estimated through interpolation. It is important to recognize the features of this process. First, the pattern of growth ascribed to the average student is arbitrarily defined to be 1 month's growth per month and is anchored in reality at only one point--the month of the school year in which the test was standardized. Second, there is no empirical information on the pattern of achievement for low-achieving (or high-achieving) students. The one empirical point is based only on the 50th percentile student. Hence, the assumption that this overall pattern holds for pupils other than the average student has little empirical basis.

A small and growing number of tests are normed on the basis of two test administrations--one in the fall and one in the spring. A procedure similar to the one described above is followed except that the grade-equivalent scale is empirically anchored at two points instead of one. The fall-to-spring interval, however, is still arbitrarily divided into equal intervals (the number of months between test points) and the spring-to-fall interval is likewise divided, again considering the 3-month summer to represent 1 month of growth. This procedure, while a little more soundly based for the average student, is still dependent upon the median student and reflects no empirical data for low-achieving students.

Despite these problems, the popular appeal of the month-for-month standard is understandable. If one believes that this is the rate of achievement for the average student and thus higher than that for the below-average student, it is reasonable to conclude that a program is effective if it produces month-for-month rates of gain for disadvantaged students.

Ten Percentile Points

The second standard that we apply is an increase of 10 percentile points from pretest to posttest. The use of a percentile point increase as a standard is based on the assumption that a student is expected to maintain the same percentile ranking from one test level to the next and

from one time to the next. Thus, in the absence of an intervention, a student who scores at the 20th percentile at the beginning of first grade would be expected henceforth to score at the 20th percentile. In other words, the test norms assume that relative rank among individuals is preserved. It should be noted that this standard is not dependent on time, as rates of gain are; that is, an increase of 10 percentile points is considered significant whether it occurs over a 3-month period or a 3-year period.

The choice of ten percentile points stems from the need to have a shift that is large enough to be educationally significant while minimizing the possibilities of chance fluctuation. Although it is impossible to determine precisely when a difference is large enough to have educational meaning, evaluators such as the RMC Research Corporation have applied a rough rule of thumb: the gain should equal or exceed one-third of the standard deviation of the norm group. We roughly estimated the equivalent of one-third of an average standard deviation by first translating one standard deviation for each test and grade level into percentile points. We then averaged these across tests and grades and arrived at 30 percentile points, one-third of which is 10 percentile points.*

Overall, the ten-percentile standard is somewhat arbitrary and extremely stringent--one which to our knowledge has never been met. For example, the final analysis of the national Follow Through evaluation data uses a standard of 1/4 standard deviation, which is not achieved in a large majority of the comparisons made. Nevertheless, for purposes of

---

*The translation varies somewhat across the distribution of test scores. For example, at the 50th percentile, an increase of one-third of a standard deviation on the CTBS roughly translates into a 13-point percentile increase compared with an 11-point percentile increase at the 20th percentile. This would pose a serious problem if we were dealing with the entire range of test scores. Because our calculations were limited to the lower portion of the distribution (centering around the 20th percentile), the problem is minimized, but not eliminated.

illustrating the impact of different standards currently in use, it serves well.

Eight Months Annual Gain

The third standard we apply is the achievement of 8 grade-equivalent months during a 12-month period. This standard, expressed in the language of test norms, is based on an expectation of 7 months annual gain for disadvantaged students. It differs from the month-for-month standard in that empirical data support this figure as an estimate of expected annual growth for disadvantaged students. One source for such support is the data collected by the states in evaluating Title I programs. If one divides each grade's mean pretest score by the number of years the students have been in school, the average annual growth is approximately 7 months across all grades (Thomas and Pelavin). The pretest scores probably include some students who were previously in Title I, suggesting that the expectation, if biased, is an overestimation. Based on this expectation, we have chosen a 1-month increase over expected achievement (that is, a total of 8 months achievement), as our third standard for judging effectiveness.*

Section III describes the type of data sought and obtained for our analyses. This is followed by a description of the analyses and the results.

---

*This 1-month increase is not related to the expected 1-month summer gain for the 50th percentile student. We have arbitrarily defined the standard to be 1 month greater than the 7-month annual expectation for the disadvantaged student. Although the 8-month standard is an annual standard, we justify applying it to the shorter fall-to-spring period in light of our hypothesis that losses occur over the summer.

III DESCRIPTION OF SEARCH FOR DATA AND DATA SETS OBTAINED

To be able to carry out comparisons between different time periods and to apply different standards, we required data with certain characteristics. Ideally, we wanted fall and spring standardized achievement test scores for individuals in raw score form for several consecutive years and several waves of students. Consecutive years of fall and spring testing permit a comparison of evaluations based on a school-year period as well as a 12-month period. Raw scores permit transformations into grade-equivalents and percentiles, thereby allowing application of the three aforementioned standards.

We restricted our search to current programs so that we could observe them in operation. Since we limited ourselves to programs whose stated objective is to increase achievement as measured by standardized tests, we required some assurance that the operating program was in fact primarily academic. We wanted to eliminate the possibility that the data might be based on programs that, in fact, did not really exist. We did not, however, pursue the issue to the point of investigating the extent to which the curricular content of the program matched the content of the test.

The remainder of this section includes a brief review of previous research on the effectiveness of compensatory education programs. In addition, we describe our search for data and the data sets obtained for analysis.

Previous Research

Our review was carried out with the idea of investigating the sustained effectiveness of compensatory programs. Therefore, we concentrated on locating research that included measures taken after the students had completed a program. Our review of the preschool literature

drew heavily on four excellent and comprehensive reviews (Stearns, 1971; White et al., 1973; Bronfenbrenner, 1974; Goodson and Hess, 1976). All the reviews indicated that substantial evidence exists to show significant short-term effects as measured primarily by standardized intelligence tests given at the end of a program. The evidence for sustained effects, based on measures taken at varying times after program participation, suggests that most cognitive gains made in preschool disappear by the second or third grade. Parent-child intervention programs are a possible exception. While these conclusions from the preschool literature are not beyond question, they at least represent a consensus of several reviewers. No such consensus exists beyond preschool.

For the early grades, Grades K-3, our review uncovered a considerable amount of research on short-term effectiveness (the references for these studies are in Appendix A). However, we were able to find virtually no work on sustained effectiveness. A study is currently under way that is designed to investigate sustained effectiveness: The Office of Education's "Study of Sustaining Effects of Compensatory Education on Basic Cognitive Skills." Preliminary results from this study are not expected before 1979, and the final results several years later.

In the remaining grades, Grades 4-12, there has again been research on short-term effectiveness (also referenced in Appendix A) but no work on sustained effectiveness. This research is not as extensive, by grade, as the research done on preschools or Grades K-3, probably because there are far fewer compensatory programs at grade levels beyond Grade 6.

Given the paucity of studies beyond preschool with measures of sustained effects, we were unable to draw from our review any conclusions about sustained effectiveness of compensatory education programs in the elementary and later grades.

Search for Data

We limited the search for data to compensatory programs beyond preschool with emphasis on the later grades. Our search for adequate data included a thorough review of projects identified in previous

searches for "exemplary" programs, an examination of ERIC and the Current Index of Education Journals, and a phone survey of large cities. Additionally, we investigated data collected as part of the evaluation of the Voucher Demonstration in the Alum Rock School District.

We devoted considerable resources to tracking down projects previously identified as "exemplary" in USOE-sponsored research done by the American Institutes for Research and the RMC Research Corporation. Since this prior research had been concerned with the quality of data, we felt the exemplary programs were our most promising source of adequate data. Of the over 40 projects reviewed, we found 15 that might have adequate data. Of these, eight were immediately eliminated when it was found they no longer existed. Six did not have data that would support reanalysis, and one program had adequate data, but obtaining it would have been prohibitively expensive.

We were quite surprised that this research did not result in the location of usable data, and that so few of the "exemplary" programs were still in existence. Because the results of this search were surprising, we have recorded the process involved and the findings in considerable detail in Appendix B.

Through our searches of ERIC and the Current Index of Education Journals, we identified two compensatory programs that might have adequate data. Although we reviewed a large number of studies, very few reported achievement test data. Most contained very general evaluation data such as teacher judgments. Of the two promising candidates, one was eliminated because the testing had not been systematic and the sample of program participants with the same tests for more than a year was extremely small. The second program was the Diagnostic-Prescriptive-Individualized Primary Reading Program in Louisville, Kentucky. We contacted the county school district and obtained permission to reanalyze their data. On receiving and attempting to analyze the data, however, we discovered limitations that precluded their use for this report.

Through phone calls to the 24 cities with the largest populations between the ages of 18 and 35, we located six metropolitan districts with potentially usable data from evaluations of compensatory education programs. Of these six districts, two had programs whose data met most of our criteria: High Intensity Learning Centers in Omaha, Nebraska and the California State Demonstration Program in Mathematics in Long Beach, California. We therefore contacted each of the programs and obtained permission to reanalyze their data.

In Long Beach, we were told about California State Demonstration Programs in other junior high schools. We contacted the Demonstration Program in Reading in Santa Barbara, California, and obtained permission to reanalyze their data.

Data from the evaluation of the Voucher Demonstration in the Alum Rock School District in San Jose, California met most of our criteria. In using these data, we recognized that increasing achievement was only one of many goals of the program, and perhaps not a primary goal. We obtained permission for our reanalysis from the Rand Corporation, which had collected the data, and from the National Institute of Education, which sponsored the demonstration.

We report on the reanalysis of data from a total of four compensatory education programs. The programs and the characteristics of the data are described below.*

Data Sets

The four data sets subjected to reanalysis represent two state-funded compensatory education programs in California, one Title I program, and the Voucher Demonstration in Alum Rock. A brief description of each program and the characteristics of the data obtained are given below.

---

*We would have liked to have detailed information on summer school participation for each program but were unable to obtain it. For three of the four programs (excluding Alum Rock for which we have no information), the program directors felt that very few students attended summer school programs but they did not have exact numbers nor individual data.

Appendix C contains some notes on the process and problems involved in obtaining and transforming the available data into a form amenable to reanalysis.

California State Demonstration Programs in Intensive Instruction in Reading and Mathematics

In 1969, the California State Legislature made funds available for the implementation of demonstration programs in reading and mathematics at the junior high school level. The intent of the legislation was to provide instructional aid to all students in about 20 junior high schools with high concentrations of educationally disadvantaged youth. The program began in Grade 7 in January 1970 and moved with the students to Grade 8 in 1970-71 and Grade 9 in 1971-72. In 1972-73, the 3-year cycle began again. Additionally, in some districts other compensatory funds were used to replicate the program in those grades not supported by the State. We obtained data from two such programs: a mathematics program in Long Beach in which district funds were used to support the program in years not funded by the State, and a reading program in Santa Barbara, which did not have district-funded replications.

Demonstration Program in Mathematics (DPM), Long Beach, California

Program Description--In Long Beach, the DPM served all students in two junior high schools, beginning in 1969-70 and 1971-72, respectively. The assumption underlying the mathematics program is that junior high school students can increase their competence in mathematics most effectively if they are given individualized instruction fitted to their needs. The program's staff have prepared a large variety of materials geared to individualized instruction including study packets designed to teach 750 behavioral objectives, criterion-referenced pretests and posttests for various skills and concepts, laboratory lessons, and review sheets.

Initially, each student is administered a criterion-referenced test to determine where in a sequence the student should begin. The program in each classroom begins each day with a Quickie Quiz, which is

a motivational technique for focusing the attention of the students. When the quiz is completed (3 to 5 minutes), one-fifth of the students go to a mathematics laboratory. Thus each student spends one day each week in the laboratory instead of the regular classroom. The laboratory lessons are designed to match the students' classroom work and are presented under the direct supervision of the lab teacher and teacher aides. At the end of the Quickie Quiz, students remaining in the classroom complete a short drill session using review sheets, and then proceed with their individual packets. This procedure is followed throughout the school year in all grades.

Data Description--Students in the program were administered the mathematics portion of the Comprehensive Test of Basic Skills (CTBS) annually in early to middle October and May. In the first year, 1969-70, the pretest was not given until January. All students received Form Q3 of the CTBS in Grade 7 and Form R3 in Grade 8. In Grade 9, the level changed to R4.* The tests were administered by counselors and members of the district evaluation staff and scored by the evaluation staff. We obtained data in raw scores for four cohorts of students. For two of the cohorts, there were data from a test given subsequent to participation in the program (administered as part of the district testing program).

Demonstration Program in Reading (DPR), Santa Barbara, California

Program Description--The reading program in Santa Barbara began in the seventh grade during the 1969-70 school year and continued with this wave of students through ninth grade. The program is in fact two separate programs: a developmental program for students considered to

*For one group of students, Cohort 4, one of three different levels of CTBS (R2, R3, or R4) was administered based on a student's preceding spring score. Additionally, the first groups of students, Cohort 1, received Form Q3 in Grade 8 and Form Q4 in Grade 9.

be average or above average and a remedial and enrichment program for the other students. Because we were mainly interested in students considered to be educationally disadvantaged, we concentrated on the remedial and enrichment component. The remedial and enrichment program was developed on the belief that learning problems are a function of environmental, academic, and psychological factors, and that students learn in different ways. Therefore, in addition to an eclectic classroom approach, the program uses the services of a staff counselor, a nurse, psychologists and home visitors.

Students identified as having reading problems spent 45 minutes daily in the Reading Complex. Those identified as having severe problems may have spent two 45-minute periods in the Reading Complex. The periods of reading are primarily individualized and small group instruction. Students' needs are identified on the basis of a variety of tests as well as information from the counselor, psychologist, or others acquainted with the students. The classes are small, 10-15 students, with a teacher, a teaching assistant, and usually a student teacher or adult volunteer, who employ a variety of instructional approaches and materials. The curriculum stresses, through reading, concepts such as cause and effect, which are taught when possible through problem-solving situations, inductive reasoning, and discovery. Also, when their schedule permits, students can attend the Reading Complex at any time in addition to their scheduled periods of participation.

Data Description--We obtained data in raw scores for one cohort of students in Santa Barbara--those starting the program in Grade 7 in 1972-73. These students were administered the reading portion of the Comprehensive Test of Basic Skills (CTBS) in October and May of each of the three years. Form Q3 of the CTBS was given in Grade 7, Form Q3 or R3 in Grade 8, and Form R4 in Grade 9. The tests were administered and scored by the program's staff.

High Intensity Learning Centers (HILINC), Omaha, Nebraska

Program Description--In 1971-72, Omaha adopted High Intensity Learning Centers (HILINC), a program developed by Random House Publishers. HILINC's purpose is to improve reading comprehension test scores of Title I students. The program serves approximately 3,500 students annually in Grades 3-12 in Title I schools. One or more High Intensity Learning Centers is at each participating school. Each center is staffed by a teacher and one teacher aide. Participating students, selected on the basis of previous test scores, spend 1 hour daily in the program in addition to their regular reading class. Initially, each student is diagnosed on the basis of an instructional objectives test. Specific materials and activities are then prescribed. These materials are intended to be self-directing and self-correcting, and are sequenced so that pupils need a minimum of teacher direction. While the materials used initially were those of the publisher, over the last 3 years the original program has been almost entirely replaced by materials written by the reading consultants and teachers.

Data Description--Omaha evaluates its Title I program on the basis of fall (early October) and spring (mid-May) administrations of the Reading Comprehension Subtest of the Gates-MacGinitie Reading Test. The level of the test is determined by a student's instructional level. Thus, for a given grade, students may receive any of several levels of the test. The tests were administered and scored by teachers in the program. We obtained scores in grade-equivalents for students in Grades 3-8 for the school years 1971-72 through 1974-75.

The Voucher Demonstration in the Alum Rock School District, San Jose, California

Program Description--In 1972 the Federal Office of Economic Opportunity (OEO) authorized a Voucher Demonstration in the Alum Rock School District. This demonstration included 6 of the district's 24 schools serving students in Grades K through 8. Each school was required to

provide at least two "mini-schools" (program options), with the district supplying the basic voucher from its current income and OEO providing compensatory vouchers for children who qualified for the Federal free lunch program. In 1973-74, the demonstration expanded to 13 schools with about 9,000 students and 45 "mini-schools," and the National Institute of Education took over sponsorship of the demonstration.

The Alum Rock Voucher Demonstration does not have one "program" in the sense of an identifiable classroom model with a specific educational goal. Instead, it reflects a large number of goals that vary somewhat from year to year. In this way it differs considerably from the other programs included in this study. There seems to be general agreement that the original intent of the demonstration has not been realized. The primary purpose of the program now seems to be to decentralize school-district authority and to provide parents with some freedom in the selection of a school program for their children. Given this purpose, it is certainly not obvious that standardized achievement tests should be the primary outcome measure, although there is clearly a consensus that one of the many goals of the demonstration is to increase cognitive achievement. This concern is discussed more fully with the presentation of the analysis results in Section IV.

Data Description--The Rand Corporation directed the testing program, which consisted of the administration of the Metropolitan Achievement Tests (MAT) in the fall and spring during the years 1972-73 through 1974-75. The tests were given in November and May of the first year and October and April of the next 2 years to students in Grades 1-8. The tests were administered by a variety of personnel including classroom teachers, members of the district's evaluation staff, and substitute teachers. The tests were scored under the auspices of the Rand Corporation. We had access to raw score data for all students tested. The data were complicated by the fact that there was no consistent pattern in the selection of alternative forms and levels of the tests. As a result, for each test point, a variety of levels and forms of the test

were administered to students in a given grade, so that a particular student often did not receive the same level of the test more than twice.

IV ANALYSIS OF PROGRAM EFFECTIVENESS

The data are presented separately by program and discussed in the following way. First we present the three means for the sample of students who were tested three times. From these means, the achievement gain and the rate of achievement are calculated for both the 9-month, school-year period and the 12-month, fall-to-fall period. The achievement gain and the rate of achievement for the school-year period are then compared with the gain and rate for the 12-month period. We then apply the three standards--a 10-point percentile increase, a gain of 8 grade-equivalent months, and an achievement rate of 1 grade-equivalent month per month--to the results for each time period. This allows us to compare the extent to which conclusions about program effectiveness vary both under different time periods and with the application of different standards.

Our discussion is extended to 2 years of a program by using samples of students who have had five tests administered to them: fall and spring of 2 consecutive years and fall of a third year. We present these five means with the achievement gains based on three different time periods: the two fall-to-spring periods, fall of the first year to spring of the second year and fall of the first year to fall of the third year. To demonstrate the extent to which the inclusion of the summer months affects an evaluation, these time periods reflect the exclusion of both summer intervals, the inclusion of the intervening summer, and the inclusion of both summers, respectively. We then consider these findings in the context of the three standards described above.

Because the standards that we apply are in terms of grade equivalents and percentiles, we report only these metrics in the text.* For data sets that contained standard scores as well, we report these scores and their standard deviations in the appendices. To simplify the text further, we present in the tables summary figures averaged across cohorts of students. In general, for all the data sets the pattern of the means for each cohort follows the pattern of the means averaged across cohorts. The data, broken down by cohort, are also presented in the appendices. References to the corresponding appendix tables appear in the text for each program.

DPM in Long Beach

From the Long Beach DPM we obtained data for four groups of students: students who began Grade 7 in 1969-70, 1970-71, 1971-72 and 1972-73. Table IV-1 presents data by grade level for all students with three test points (fall and spring of one year and fall of the next)† averaged across four groups. The first three columns span a 12-month period and contain the grade-equivalent and the percentile scores associated with each standard score mean for each test administered for Grades 7 and 8. These statistics, as well as the standard score means and standard deviations, are presented separately by grade, school, and cohort in Appendix D. We are primarily interested in comparing the achievement over the traditional fall-to-spring evaluation period with

---

*With the exception of the Omaha program, which reported only grade-equivalent scores, the means were always calculated in standard scores and then translated into grade-equivalents. This avoids the problems associated with averaging grade-equivalents. For all the data sets, we compared calculations based on means and medians and found no difference in the resulting patterns.

†To determine if our samples are representative of all students in the program, we have compared our samples to all students tested at a given point. We have found no systematic differences between the means and standard deviations of our samples and the larger, cross-sectional groups. In general, where there are differences they tend to favor the longitudinal groups, which is not surprising since they probably represent a more stable group. The cross-sectional data are presented in Appendix D.

Table IV-1

LONG BEACH DPM CTBS MATHEMATICS MEANS IN GRADE-EQUIVALENTS AND PERCENTILES AND GAINS FOR TWO TIME PERIODS

Sample	Means I Fall	Means II Spring	Means III Fall	Gains IV Fall to Spring	Gains V Fall to Fall
Grade 7 (n=780)					
Grade-equivalent	5.5	7.4	6.6	1.9	1.1
Percentile	23	45	28	22	5
Grade 8 (n=468)					
Grade-equivalent	6.4	7.9	7.8	1.5	1.4
Percentile	26	38	30	12	4

the achievement over the 12-month, fall-to-fall period. Comparing the grade-equivalent and percentile means in Columns II and III, the second fall score is lower than the spring score for both grades. Therefore, the fall-to-fall estimates of achievement (Column V) are smaller than the fall-to-spring estimates (Column IV). The small difference in grade equivalents for Grade 8 reflects the small difference in the means. Since the level of the test changed between the spring of Grade 8 and the fall of Grade 9, the smaller summer loss for the Grade 8 samples may be a function of the level change. Since the test level change is completely confounded with program participation in Grade 8, it is impossible to be certain of the cause.

We now consider the impact of these summer losses on conclusions about program effectiveness as judged by the three standards described above. First, we inspect shifts in percentile scores under the assumption that they would remain the same, on the average, in the absence of a program impact. We then compare increases in percentile to our most stringent standard, that of a 10-point increase for the two time periods, fall to spring and fall to fall. Looking at the percentile differences for the fall-to-spring period in Column IV, we see that there is a substantial percentile increase for both grades: 22 and 12 percentile points. Both of these increases exceed the 10-point standard. However,

if one looks at the percentile changes for the fall-to-fall period, Column V, a very different picture emerges. Here the percentile increases are only 5 and 4 points. For this time period, neither grade reaches the standard of a 10-point increase. Hence, while the program would be judged quite effective for a fall-to-spring period, it would not be judged so for a fall-to-fall period.

Our second standard is a gain of 8 grade-equivalent months (0.8 grade-equivalent years) per year. If we look at the grade-equivalent gains in Column IV for the fall-to-spring period (less than a year), the program looks extremely effective. The Grade 7 gain is 19 months and the Grade 8 gain is 15 months. When the summer is included in the period over which the gain is measured, however, the gains are much smaller (see Column V). Nevertheless, while the gains are smaller for the fall-to-fall period (11 and 14 months), both grades still exceed the standard of an 8-month gain per year. Hence, the program would still be considered effective.

The third standard is a gain of 1 grade-equivalent month per month. Table IV-2 gives the average monthly grade-equivalent rate for the fall-to-spring (Column I) and fall-to-fall (Column II) periods. These rates are calculated by dividing the fall-to-spring and fall-to-fall gains from the totals in Table D-3* by 7 and 10 respectively.†

Comparing Column I with Column II in Table IV-2, we see that again the rates are substantially smaller for the fall-to-fall period. If the program is judged on the basis of the fall-to-spring rates, it is quite effective, with monthly rates of 2.8 and 2.1 months per month. However, these rates diminish considerably when calculated over the fall-to-fall

*The appendix tables provide the rates averaged across cohorts. Therefore the rates are slightly different than those calculated directly from Table IV-1 due to rounding error.

†The divisor for the fall-to-spring period is 7 since the interval between the fall and spring administrations of the CTBS is 7 months. The divisor for the annual rate is 10 since the grade-equivalent year contains 10 grade-equivalent months, 9 for the school year and one for the summer.

Table IV-2

LONG BEACH DPM CTBS MATHEMATICS MONTHLY ACHIEVEMENT RATES IN GRADE-EQUIVALENT MONTHS FOR TWO TIME PERIODS

Sample	Monthly Achievement Rates I Fall to Spring	II Fall to Fall
Grade 7 (n=780)	2.8	1.2
Grade 8 (n=468)	2.1	1.3

period. For this time period they are 1.2 and 1.3 months per month. Nevertheless, even for the fall-to-fall period, the program overall is still effective, with both rates in excess of 1 month per month.

No matter which standard is applied, we argue that fall to fall is the appropriate period of time for judging program effectiveness. If an evaluation is based on a traditional fall-to-spring period, the results will not reflect the extent to which gains have lasted, at least until the beginning of the next school year. The Long Beach data illustrate that for 1 year, the fall-to-fall gains are consistently smaller than fall-to-spring gains. However, the gains are sufficiently large during the school year that, in spite of large summer losses, the program is judged effective under two of the three standards of effectiveness over the 12-month, fall-to-fall period.

We now extend our analysis to judgements of 2 years of the program with a sample of students who were tested five times: fall and spring of Grades 7 and 8 and fall of Grade 9. For each of the five test administrations Table IV-3 presents the grade-equivalent mean and the percentile associated with each standard score mean. Appendix D presents these data, as well as the standard score means and standard deviations, separately by cohort and school.

This 2-year sample reflects the same pattern as the two 1-year samples described above. There are losses over both summer intervals (Column III minus Column II and Column V minus Column IV). Again, it

Table IV-3

LONG BEACH DPM CTBS MATHEMATICS MEANS IN GRADE EQUIVALENTS AND PERCENTILES FOR TWO YEARS

	Means				
	I	II	III	IV	V
Sample	Grade 7 Fall	Grade 7 Spring	Grade 8 Fall	Grade 8 Spring	Grade 9 Fall
Grades 7-8 (n=378)					
Grade-equivalent	5.5	7.3	6.6	8.1	7.9
Percentile	23	43	28	40	31

should be noted that the difference over the second summer reflects a change in test level, which may or may not explain the smaller loss.

Since there are losses over both summers, 2-year estimates of achievement will be largest if neither summer is included; that is, if 2-year achievement is measured as the sum of two fall-to-spring gains. This time period yields a gain of 1.8 years (Column II minus Column I) plus 1.5 years (Column IV minus Column III) which is a gain of 3.3 grade-equivalent years for the 2 years. If the estimate of 2-year achievement includes the intervening summer, the estimate of achievement is lowered to 2.6 grade-equivalent years (Column IV minus Column I). Finally, if both summers are included, the achievement estimate is even smaller--2.4 grade-equivalent years (Column V minus Column I). Similarly, in the percentile metric, the sum of the two fall-to-spring gains is 32 percentile* points. Inclusion of the intervening summer reduces the gain to 17 percentile points, and the inclusion of both summers lowers the gain to 8 percentile points.

In comparing the differences under the three time periods to the 10-point percentile standard, it is obvious that the 32 percentile

---

*If the Grade 8 fall score reflects any part of the impact of the Grade 7 programs, creating the two fall-to-spring gains separately is misleading. Logically, the Grade 7 fall score should serve as the expected percentile throughout the program.

point gain calculated by summing the two fall-to-spring gains greatly exceeds the 10-point standard. The percentile shift is smaller (17 points) when measuring from the fall of Grade 7 to the spring of Grade 8, but still large enough to meet the standard for program effectiveness. However, when both summers are included in the evaluation by measuring from the fall of Grade 7 to the fall of Grade 9, the percentile increase of 8 points no longer reaches the standard. Hence, under the time period measuring sustained effects, the program would not be judged effective.

We now compare the grade-equivalent gains to the standard of an 8 grade-equivalent month gain per year in order to evaluate the success of the program. This means that the effectiveness of a 2-year program should be judged by comparing the 24-month gain (fall of Grade 7 to fall of Grade 9) to 1.6 grade-equivalent years (a gain of 0.8 year or 8 months for each year). For all three time periods the program is effective using this standard. While the inclusion of both summers gives the smallest gains, the fall of Grade 7 to the fall of Grade 9 still reflects a gain of 2.4 grade-equivalent years, which exceeds the standard of 1.6 grade-equivalent years.

Turning to Table IV-4, we now compare the 2-year rates of growth in grade-equivalent months to the standard of a month-for-month gain. The first two columns present the monthly rates based on the two fall-to-spring intervals. These rates are 2.8 and 2.0 months per month, respectively.* If the program were judged on this basis it would be considered effective over a 2-year period by virtue of greatly exceeding the standard in both years. If the program were judged on a time frame including the intervening summer, the rate of 1.6 months per month still exceeds the month-for-month standard. Finally, judged on the full 2-year time period (Column IV), the rate of 1.3 months per month is even smaller but still exceeds the standard.

---

*These rates are based on the totals in Appendix Table D-5, which are the rates averaged across cohorts. Therefore, the rates are slightly different than they would have been if calculated directly from Table IV-3, due to rounding error.

Table IV-4

LONG BEACH DPM CTBS MATHEMATICS MONTHLY ACHIEVEMENT RATES IN GRADE-EQUIVALENT MONTHS FOR THREE TIME PERIODS

	Monthly Achievement Rates			
Sample	I Grade 7 Fall to Spring	II Grade 8 Fall to Spring	III Grade 7 Fall to Grade 8 Spring	IV Grade 7 Fall to Grade 9 Fall
Grades 7-8 (n=378)	2.8	2.0	1.6	1.3

In summary, the Long Beach data illustrate that estimates of achievement and effectiveness can vary tremendously when different time frames are used in both 1-year and 2-year evaluations. While the Long Beach program continues to look effective under all time periods for the two grade-equivalent standards, it is important to keep in mind that the inclusion of the summer months does reduce the size of the achievement gains and, in the case of the 10-point percentile standard, changes the conclusions reached.

DPR in Santa Barbara

For the Santa Barbara reading program, we have data for only one cohort of students, those who entered Grade 7 in 1972-73. Columns I, II, and III in Table IV-5 contain the grade-equivalent and the percentile associated with each mean for three test administrations for all students tested in fall and spring of one year and fall of the next.* The standard score means and standard deviations are presented in Appendix E. Columns IV and V give the gains from fall to spring and fall to fall, respectively. For both grades, there is a loss of achievement during the summer. This summer loss is reflected in the comparison between the

*We then compared the means and standard deviations of these samples to all students tested at each test point and found no differences. See Appendix E.

Table IV-5

SANTA BARBARA DPR CTBS READING MEANS IN GRADE-EQUIVALENTS AND PERCENTILES AND GAINS FOR TWO TIME PERIODS

	Means			Gains	
	I	II	III	IV	V
Sample	Fall	Spring	Fall	Fall to Spring	Fall to Fall
Grade 7 (n=102)					
Grade-equivalent	4.3	5.6	5.4	1.3	1.1
Percentile	12	20	16	8	4
Grade 8 (n=107)					
Grade-equivalent	5.5	6.5	6.2	1.0	0.7
Percentile	16	23	16	7	0

fall-to-spring and fall-to-fall gains. The fall-to-fall estimate of achievement is smaller than the fall-to-spring estimate by 2 grade-equivalent months in Grade 7 and 3 in Grade 8. This difference is also reflected in percentile shifts, where the gains are 8 and 7 percentile points for the two grades as measured from fall-to-spring, but only 4 and 0 points for the two grades when measured from fall-to-fall.

A comparison of these percentile shifts to the 10-point standard shows that the program does not meet the standard under either time period. However, a comparison of the grade-equivalent gains to the standard of an 8-month gain per year shows that the program is effective in both grades from fall to spring. During the 12-month period, the program is effective in Grade 7 (a fall-to-fall gain of 1.1 years or 11 months) but not effective in Grade 8 (a fall-to-fall gain of 0.7 year or 7 months).

Table IV-6 presents the monthly rates for the two time periods. Comparing these with the month-for-month standard, we see that for both grades the fall-to-spring rates exceed the standard (1.9 and 1.4). However, during the fall-to-fall period, the incorporation of the summer into the estimate lowers the rates to 1.1 and 0.7 month per month--only the Grade 7 program is effective under the month-to-month standard.

Table IV-6

SANTA BARBARA DPR CTBS READING MONTHLY ACHIEVEMENT RATES IN GRADE-EQUIVALENT MONTHS OVER TWO TIME PERIODS

	Monthly Achievement Rates	
	I	II
Sample	Fall to Spring	Fall to Fall
Grade 7 (n=102)	1.9	1.1
Grade 8 (n=107)	1.4	0.7

Table IV-7 extends the data to 2 years of the program with means for students with five consecutive test points (fall and spring of Grades 7 and 8 and fall of Grade 9). Again, there is a loss during both summers, 1 grade-equivalent month or 4 percentile points over the first summer (Column III minus Column II) and 3 grade-equivalent months or 6 percentile points over the second summer (Column V minus Column IV). Consequently, the inclusion of each summer in the evaluation time period reduces the size of the achievement gain.

We first compare the changes in percentile scores under th. three periods to the standard of a gain of 10 percentile points. The sum of the two fall-to-spring gains is 14 percentile points, which clearly

Table IV-7

SANTA BARBARA DPR CTBS READING MEANS IN GRADE-EQUIVALENTS AND PERCENTILES FOR TWO YEARS

	Means				
	I	II	III	IV	V
Sample	Grade 7 Fall	Grade 7 Spring	Grade 8 Fall	Grade 8 Spring	Grade 9 Fall
Grade 7-8 (n=99)					
Grade-equivalents	4.3	5.6	5.5	6.5	6.2
Percentile	12	20	16	22	16

exceeds the 10-percentile point standard.* Using only the fall-to-spring gains the program would be judged effective.

The increase from the fall of Grade 7 to the spring of Grade 8 is 10 percentile points. However, if the program is judged on the basis of sustained gains over both summers, and measured from the fall of Grade 7 to the fall of Grade 9, the increase in percentiles is only 4 points. During this time period, the program would not be judged effective.

We next compare the grade equivalent gains during the three time periods to the standard of an 8-month gain during each year or a 16-month gain during 2 years For all three time periods, the program is judged effective when using this standard. The sum of the two fall to spring gains is 23 months; the gain from the first fall to the second spring is 22 months; and the gain from the first to third fall is 19 months. Each gain is greater than the 16-month standard.

Table IV-8 contains the monthly rates in grade-equivalents for the three 2-year time periods. Under all three time periods the program is judged effective when compared with the standard of a month-for-month gain. However, this rate is considerably smaller (1.0) when measured over the full two calendar years than when measured by excluding one summer (1.3) or both summers (1.9 and 1.4 for the two school years).

Table IV-8

SANTA BARBARA DPR CTBS READING MONTHLY ACHIEVEMENT RATES IN GRADE-EQUIVALENT MONTHS FOR THREE TIME PERIODS

	Monthly Achievement Rates			
Sample	I Grade 7 Fall to Spring	II Grade 8 Fall to Spring	III Grade 7 Fall to Grade 8 Spring	IV Grade 7 Fall to Grade 9 Fall
Grades 7-8 (n=99)	1.9	1.4	1.3	1.0

*If the Grade 8 fall score reflects any part of the impact of the Grade 7 programs, creating the two fall-to-spring gains separately is misleading. Logically, the Grade 7 fall score should serve as the expected percentile throughout the program.

These findings again illustrate that results vary over different periods of time, and that such differences can affect conclusions about program effectiveness.

Omaha HILINC

From the HILINC program in Omaha, we obtained data for six different cohorts of students in both public and nonpublic schools spanning Grades 3-8 for a 4-year period. We present only the data from students in public schools averaged across cohorts. Data for public school students are presented by cohort in Appendix F. Data for students in nonpublic schools are presented in Appendix G. Analyses were performed only in those grades and cohorts for which there were at least 20 students for whom we had received data. Since Omaha records test results only in grade-equivalents, our analyses were restricted to this metric.

Table IV-9 contains the grade-equivalent means for all students with at least three test points (fall and spring of one year and the fall of the next year).* We first compare the means for the spring

Table IV-9

OMAHA HILINC GATES-MacGINITIE READING MEANS IN GRADE-EQUIVALENTS AND GAINS FOR TWO TIME PERIODS

	Means			Gains	
Sample	I Fall	II Spring	III Fall	IV Fall to Spring	V Fall to Fall
Grade 3 (n=272)	2.2	3.3	2.8	1.1	0.6
Grade 4 (n=931)	2.6	3.6	3.2	1.0	0.6
Grade 5 (n=980)	3.3	4.3	4.0	1.0	0.7
Grade 6 (n=316)	3.8	4.8	4.4	1.0	0.6
Grade 7 (n=[illegible]28)	4.3	5.2	4.9	0.9	0.6

*The corresponding data for all students tested at each point are presented in Appendix F. While the cross-sectional means are consistently higher than the longitudinal samples, the differences are extremely small.

(Column II) with the means for the following fall (Column III). For all five grades, the fall means are lower than the means of the previous spring. Consequently, for all grades achievement as measured from fall to fall is smaller than achievement from fall to spring. The differences between the estimates for the two periods of time (Column V minus Column IV) range from 3 to 5 grade-equivalent months.

Since percentiles are not available, we cannot apply the percentile standard; therefore, we turn to the two grade-equivalent standards for assessing program success. Using the standard of an 8-month gain, all the grades exceed the standard during the school year. However, for all grades the inclusion of the summer loss reduces this gain to less than 8 months. Thus, in every grade, the program would be considered effective if judged from fall to spring, but failing if judged from fall to fall.

Table IV-10 translates the fall-to-spring and fall-to-fall achievement into monthly rates by dividing the achievement by 7 months (the number of months between the test administrations) and 10 months, respectively. Comparing these rates to the month-for-month standard, we see that for all grades the monthly rate as calculated from fall-to-spring exceeds the standard. These achievement rates range from a low of 1.3 to a high of 1.6 months per month. But if we judge the program

Table IV-10

OMAHA HILINC GATES-MacGINITIE READING MONTHLY ACHIEVEMENT RATES IN GRADE-EQUIVALENT MONTHS FOR TWO TIME PERIODS

	Monthly Achievement Rates	
	I	II
Sample	Fall to Spring	Fall to Fall
Grade 3 (n=272)	1.6	0.6
Grade 4 (n=931)	1.4	0.6
Grade 5 (n=980)	1.4	0.7
Grade 6 (n=316)	1.4	0.6
Grade 7 (n=128)	1.3	0.6

on the basis of rates calculated for the 12-month year, the rates for all of the grades are below the standard. These achievement rates range from 0.6 to 0.7 month per month.

Table IV-11 presents the grade-equivalent means for those students tested at least five times (fall and spring of two successive years and fall of the next). These means are presented by grade range. The data are presented by cohort in Appendix F. All the grades show losses for both summers, ranging from 2 to 6 grade-equivalent months for the first summer and 4 to 6 months for the second summer. Thus, for all samples, the inclusion of the first summer in estimating achievement (Column IV minus Column I) will reduce the estimate from that based on the two school years. And the inclusion of both summers (Column V minus Column I) reduces the estimate of achievement still further.

We now compare the gains over the three time periods to the 8-month standard. Since we are viewing 2 years of the program, the standard for effectiveness is a gain of 1.6 years. The sum of the gains for both years based on the two fall-to-spring periods (Column II minus Column I, and Column IV minus Column III) exceeds 1.6 years in all samples. The sums range from 1.7 years to 2.1 years. Therefore, the program would be judged effective. When the gains from the first fall to the second spring (Column IV minus Column I) are used, only two of the samples reach the 1.6-year standard (Grades 4-5 and 5-6). The other two samples

Table IV-11

OMAHA HILING GATES-MacGINITIE READING MEANS IN GRADE-EQUIVALENTS FOR TWO YEARS

Sample	Means I Fall	II Spring	III Fall	IV Spring	V Fall
Grades 3-4 (n=87)	2.6	3.6	3.0	4.1	3.5
Grades 4-5 (n=324)	[illegible].7	3.5	3.1	4.3	3.9
Grades 5-6 (n=130)	3.2	4.1	3.9	4.8	4.4
Grades 6-7 (n=45)	4.1	4.9	4.5	5.4	5.0

are close, having gains of 1.5 and 1.3 years, but would not be judged effective by the 1.6-year standard. If one includes both summer intervals in order to reflect sustained achievement, the gains (Column V minus Column I) range from 0.9 to 1.2 years. Under this time period, none of the samples reaches the 1.6-year standard.

To judge the program against the standard of month-for-month achievement, we present the achievement rates for the three time periods in Table IV-12. If we compare the two fall-to-spring rates (Columns I and II) with the standard, we see that in all cases the school-year rates exceed the standard for both years. In fact, most of the rates are substantially greater than the month-for-month standard. If we include the intervening summer in estimating the achievement rate (Column III), none of the samples reaches the standard. If we now include both summers in order to capture the extent to which achievement is sustained we find another substantial drop (Column IV). The rates based on the period from the first to the third fall range from 0.5 to 0.6 month per month.

These findings provide a dramatic illustration of how conclusions about program effectiveness change when the evaluation time period includes the summer months. This program is consistently effective during the school year but, because of large summer losses, cannot be judged effective for longer periods of time.

Table IV-12

OMAHA HILINC GATES-MacGINITIE READING MEANS IN RATES IN GRADE-EQUIVALENT MONTHS FOR THREE TIME PERIODS

	Monthly Achievement Rates			
Sample	I Fall to Spring Year 1	II Fall to Spring Year 2	III Fall 1 to Spring 2	IV Fall 1 to Fall 3
Grades 3-4 (n=87)	1.4	1.6	0.9	0.5
Grades 4-5 (n=324)	1.1	1.7	0.9	0.6
Grades 5-6 (n=130)	1.3	1.3	0.9	0.6
Grades 6- (n=45)	1.1	1.3	0.8	0.5

Alum Rock Voucher Demonstration

For Alum Rock, we obtained 3 years of data including six cohorts of students in Grades 1 through 7. The reader should be reminded before inspecting the results that this is the one program that is not specifically a reading or mathematics program intended to increase scores on standardized tests. Therefore, although the numbers are interpreted in the context of program effectiveness, conclusions should be drawn with caution.

Table IV-13 contains data for all students with three test points by grade. Columns I, II, and III contain the grade-equivalent score associated with the mean for each test administration. These data along with the standard score means and standard deviations are presented in Appendix H. Columns IV and V contain the differences in grade-equivalent means for the fall-to-spring and fall-to-fall periods, respectively. For all grades, the means are based on at least two different levels of the Metropolitan Achievement Test.* Therefore, the interpretation of any one mean presumes the adequacy of the standard score scale and grade-equivalent scale across test levels.†

The most striking feature of Table IV-13 is the remarkable similarity between the spring and subsequent fall scores, and hence between the fall-to-spring and fall-to-fall achievement. Across grades, the largest difference in gains for the two time periods is 2 months for Grade 7.

A possible explanation for this finding is that the Voucher Demonstration is more an organizational scheme for schools than a program aimed specifically at reading instruction. Therefore, these scores might present a picture of untreated disadvantaged students. Without other data on untreated students, i is impossible to draw this conclusion

---

*Because the means include scores from out-of-level tests (levels not normed for that grade), percentile scores are inappropriate and therefore not included.

†Two studies of the MAT standard score scale have recently been completed (Barker and Pelavin; Pelavin and Barker). Both studies indicate that the scale may contain biases.

Table IV-13

ALUM ROCK VOUCHER DEMONSTRATION MAT READING MEANS IN GRADE-EQUIVALENT AND GAINS FOR TWO TIME PERIODS

Sample	Means I Fall	Means II Spring	Means III Fall	Gains IV Fall to Spring	Gains V Fall to Fall
Grade 1 (n=665)	1.3	1.7	1.7	0.4	0.4
Grade 2 (n=582)	1.8	2.4	2.5	0.6	0.7
Grade 3 (n=781)	2.5	3.1	3.1	0.6	0.6
Grade 4 (n=832)	2.9	3.5	3.5	0.6	0.6
Grade 5 (n=842)	3.6	4.2	4.2	0.6	0.6
Grade 6 (n=728)	4.3	4.8	4.9	0.5	0.6
Grade 7 (n=813)	5.3	6.0	6.2	0.7	0.9

with confidence. It is interesting to note, however, that these data reflect much smaller school year gains and smaller relative summer losses than those found in the three programs investigated above.

We now compare the gains and rates over two time periods to the grade-equivalent standards. Neither the fall-to-spring nor the fall-to-fall achievement gains meet the standard of an 8-month gain except for Grade 7 fall to fall. All other gains for both the fall-to-spring period and the fall-to-fall period range from 0.4 to 0.7 month.

Table IV-14 contains the monthly achievement rates for the samples with three tests. These are calculated by dividing the achievement by 6 and 10 months respectively (the number of months between the test administrations).* A comparison of the monthly achievement rates over the two time periods to the month-for-month standard makes the differences over the two time periods more pronounced. Of the seven grades, four

*The rates are based on the totals in Appendix Table H-3, which are the rates averaged across cohorts. Hence the rates are slightly different from those that would have been calculated directly from Table IV-13, due to rounding error.

Table IV-14

ALUM ROCK VOUCHER DEMONSTRATION MAT READING MONTHLY ACHIEVEMENT RATES IN GRADE-EQUIVALENT MONTHS FOR TWO TIME PERIODS

	Monthly Achievement Rates	
	I	II
Sample	Fall to Spring	Fall to Fall
Grade 1 (n=665)	0.6	0.4
Grade 2 (n=582)	0.8	0.6
Grade 3 (n=781)	1.0	0.7
Grade 4 (n=832)	1.0	0.7
Grade 5 (n=842)	1.0	0.5
Grade 6 (n=728)	0.8	0.7
Grade 7 (n=813)	1.2	0.9

reach or exceed the month-for-month standard during the fall-to-spring period. During the fall-to-fall period, the range of rates is only 0.4 to 0.9 month per month; none of the samples reaches the standard.

For students with scores for all five test administrations, Table IV-15 presents the grade-equivalents associated with each of the five standard score means. The pattern seen above for annual growth is also reflected in this 2-year sample. The differences between the spring and fall scores for both years (Column III minus Column II and Column V minus Column IV) are very small; in fact, there is no difference in 6 of the 12 comparisons. The largest difference is an increase from spring to fall of 3 months (Grades 5-6, second summer). Consequently, comparisons of achievement over the different time periods show little difference.

We first compare the achievement gains under three periods of time to the standard of a 16-month gain. When the two fall-to-spring gains are summed (Column II minus Column I plus Column IV minus Column III), only one of the six samples, that for Grades 6-7, reaches the standard. Overall, the range of the sum of the two fall-to-spring gains is 9 to 16 grade-equivalent months. When the gain is calculated from the initial fall to the second spring (Column IV minus Column I), again only one of

Table IV-15

ALUM ROCK VOUCHER DEMONSTRATION MAT READING MEANS
IN GRADE-EQUIVALENTS FOR TWO YEARS

Sample	Means I Fall 1	II Spring 1	III Fall 2	IV Spring 2	V Fall 3
Grades 1-2 (n=147)	1.4	1.8	1.8	2.3	2.5
Grades 2-3 (n=147)	1.9	2.5	2.5	3.2	3.2
Grades 3-4 (n=193)	2.5	3.1	3.2	3.7	3.7
Grades 4-5 (n=194)	2.9	3.5	3.6	4.3	4.2
Grades 5-6 (n=191)	3.7	4.2	4.2	4.6	4.9
Grades 6-7 (n=136)	4.6	5.3	5.3	6.2	6.0

the six samples (Grades 6-7) reaches the standard. The gains for this period range from 9 months to 16 months. Under the fall-to-fall period (Column V minus Column I), none of the samples reaches the standard. Here the gains range from 11 months to 14 months.

Table IV-16 contains the monthly achievement rates for three different time periods based on the means in Table IV-15. When the month-for-month standard is used, three of the six samples reach this standard during <u>both</u> of the fall-to-spring periods (Columns I and II). When these rates are calculated from the fall of the first year to the spring of the second year, only one of the samples (Grades 6-7) reaches the standard. Under the period from the fall of the first year to the fall of the third year, none of the samples reaches the standard. These rates range from 0.6 to 0.7 month per month.

Table IV-16

ALUM ROCK VOUCHER DEMONSTRATION MAT READING MONTHLY ACHIEVEMENT RATES IN GRADE-EQUIVALENT MONTHS FOR THREE TIME PERIODS

Sample	Monthly Achievement Rates			
	I Year 1 Fall to Spring	II Year 2 Fall to Spring	III Fall 1 to Spring 2	IV Fall 1 to Fall 3
Grades 1-2 (n=147)	0.7	0.8	0.6	0.6
Grades 2-3 (n=147)	1.0	1.2	0.9	0.7
Grades 3-4 (n=193)	1.0	0.8	0.8	0.6
Grades 4-5 (n=194)	1.0	1.2	0.9	0.7
Grades 5-6 (n=191)	0.8	0.7	0.6	0.6
Grades 6-7 (n=136)	1.2	1.5	1.0	0.7

Conclusions

In Section II we argued that the goal of increasing achievement of participants in compensatory education programs implies that an increase in achievement should persist beyond the end of the program. If a program does increase achievement, it is perhaps unrealistic to expect all of that increase to be maintained year after year. However, it does seem reasonable to expect part of the increase to be sustained at least through the summer following the program. If this does not occur, those concerned with compensatory education programs should have this information. Therefore, we believe that program evaluations, and the accompanying conclusions about program effectiveness, should be based minimally on a fall-to-fall time period instead of the usual fall-to-spring time period.

Unfortunately, the extent to which achievement is sustained is rarely studied, hence little data exist that speak to the issue. In this section, we have presented four sets of data that permit comparisons of achievement and effectiveness over both a fall-to-spring and a fall-to-fall period. We have made this comparison in several ways including the application of three standards of success to the results under the two

time periods. While the standards are somewhat arbitrary (in the absence of accurate information on "normal" growth for educationally disadvantaged students), these standards serve to illustrate how conclusions about program effectiveness can change under the different time periods. We also extended the analysis to 2 years of a program and carried out analogous

Table IV-17

ONE-YEAR EFFECTIVENESS BY PROGRAM BY GRADE AS JUDGED AGAINST THREE STANDARDS FOR TWO TIME PERIODS

Program and Grade	10 Percentile Point Standard: Fall to Spring	10 Percentile Point Standard: Fall to Fall	8-Month Standard: Fall to Spring	8-Month Standard: Fall to Fall	Month-per-Month Standard: Fall to Spring	Month-per-Month Standard: Fall to Fall
Long Beach DPM						
Grade 7 (n=780)	22*	5	19*	11*	2.8*	1.2*
Grade 8 (n=468)	12*	4	15*	14*	2.1*	1.3*
Santa Barbara DPR						
Grade 7 (n=102)	8	4	13*	11*	1.9*	1.1*
Grade 8 (n=108)	7	0	10*	7	1.4*	0.7
Omaha HILINC						
Grade 3 (n=272)	NA†	NA†	11*	5	1.5*	0.5
Grade 4 (n=931)	↓	↓	9*	5	1.3*	0.5
Grade 5 (n=980)	↓	↓	10*	8*	1.5*	0.8
Grade 6 (n=316)	↓	↓	9*	6	1.3*	0.6
Grade 7 (n=128)	↓	↓	9*	6	1.3*	0.6
Alum Rock Voucher Demonstration						
Grade 1 (n=665)	↓	↓	4	4	0.6	0.4
Grade 2 (n=582)	↓	↓	6	7	0.8	0.7
Grade 3 (n=781)	↓	↓	6	6	1.0*	0.7
Grade 4 (n=1111)	↓	↓	6	6	1.0*	0.7
Grade 5 (n=842)	↓	↓	6	6	1.0*	0.5
Grade 6 (n=728)	↓	↓	5	6	0.8	0.7
Grade 7 (n=813)	NA	NA	7	9*	1.2*	0.9

*The standard has been reached or exceeded.

†NA = not applicable.

comparisons. Table IV-17 summarizes the results from the four programs in terms of the three standards applied: a shift of 10 percentile points, an annual achievement rate of 8 grade-equivalent months, and a rate of 1 grade-equivalent month per month. Under each standard the results for each grade level for each program are presented, first based on a fall-to-spring period and then on a fall-to-fall period. The asterisks indicate that the standard was reached.

We have demonstrated that the fall-to-fall estimates of achievement are consistently, and often substantially, lower than the fall-to-spring estimates. This reflects the findings that large mean gains over the school year are often followed by large losses over the following summer. Hence conclusions about program effectiveness can be completely reversed when the summer interval is included in the evaluation time period. Conclusions are not always reversed, however. We have presented examples of programs that do show a sustained impact. Regardless of the conclusions reached, it is important to know if a program has a lasting impact, and thus we conclude that evaluations should be based on a fall-to-fall period instead of the traditional fall-to-spring period.

V INDIVIDUAL-LEVEL ANALYSES

The analyses in Section IV have rested primarily on inspection of means--scores averaged over individuals. Making recommendations for evaluation practices on the basis of mean-level analyses assumes implicitly that the pattern of the means is reflected in at least a majority of individual cases. If the summer-loss phenomenon occurred because a small proportion of students in each sample had enormous summer losses, rather than because most students showed losses, we would hesitate to argue strongly for changes in evaluation practices. Therefore, we have conducted a small number of individual-level analyses to determine whether the pattern of the means is reflected by individuals. The analyses are limited by time and cost constraints. We discuss first the proportion of students in five samples that show losses in achievement over the summer. We then discuss the relationships between amount of school-year gain and amount of summer loss. Finally we discuss the relationships between amount of school-year gain and amount of 12-month, fall-to-fall gain.

Proportion of Students With Summer Loss

To investigate the extent to which the summer losses shown in the mean test scores accurately reflect the patterns of individuals in the samples, we have plotted the school-year gains against the summer gains (or losses) for five samples. These samples are two cohorts of Grade 7 students in School A in Long Beach DPM, the Grade 7 sample from Santa Barbara DPR, and one cohort for each of Grades 3 and 4 in the Omaha HILINC program. As a reminder of the mean patterns found, Table V-1 presents the means for three test points (fall, spring, and fall) followed by the fall-to-spring (school-year) gains and the spring-to-fall (summer) losses for each of the five samples. The Long Beach and

Table V-1

MEANS AND DIFFERENCES FOR STUDENTS IN THE FIVE SAMPLES
SHOWN IN FIGURES V-1 TO V-5

Sample	Means			Gains	
	Fall	Spring	Fall	Fall to Spring	Spring to Fall
Long Beach DPM*					
Cohort 3					
Grade 7 (n=109)	413	459	439	46	-20
Cohort 4					
Grade 7 (n=82)	422	495	463	73	-32
Santa Barbara DPR*					
Grade 7 (n=102)	405	453	446	48	-7
Omaha HILINC†					
Cohort 1					
Grade 3 (n=152)	1.97	3.05	2.60	1.08	-.45
Cohort 2					
Grade 4 (n=387)	2.54	3.60	3.16	1.06	-.44

*Standard scores, CTBS.

†Grade-equivalents, Gates-MacGinitie.

Santa Barbara samples are presented in standard scores* (CTBS) and the Omaha scores in grade-equivalents (Gates-MacGinitie).

Figures V-1 to V-5 contain scatterplots of individual scores, school-year gain against summer gain (loss) for each of the five samples. The horizontal line drawn on each chart is the zero line indicating no gain or loss over the summer. All students whose scores fall below that line experience at least some loss in achievement over the summer. The vertical line represents zero gain over the school year. Hence, students

*Our preference is to use standard scores whenever possible since this is the only metric which is defined to be equal-interval; that is, the distance between any two adjacent points on the scale is the same.

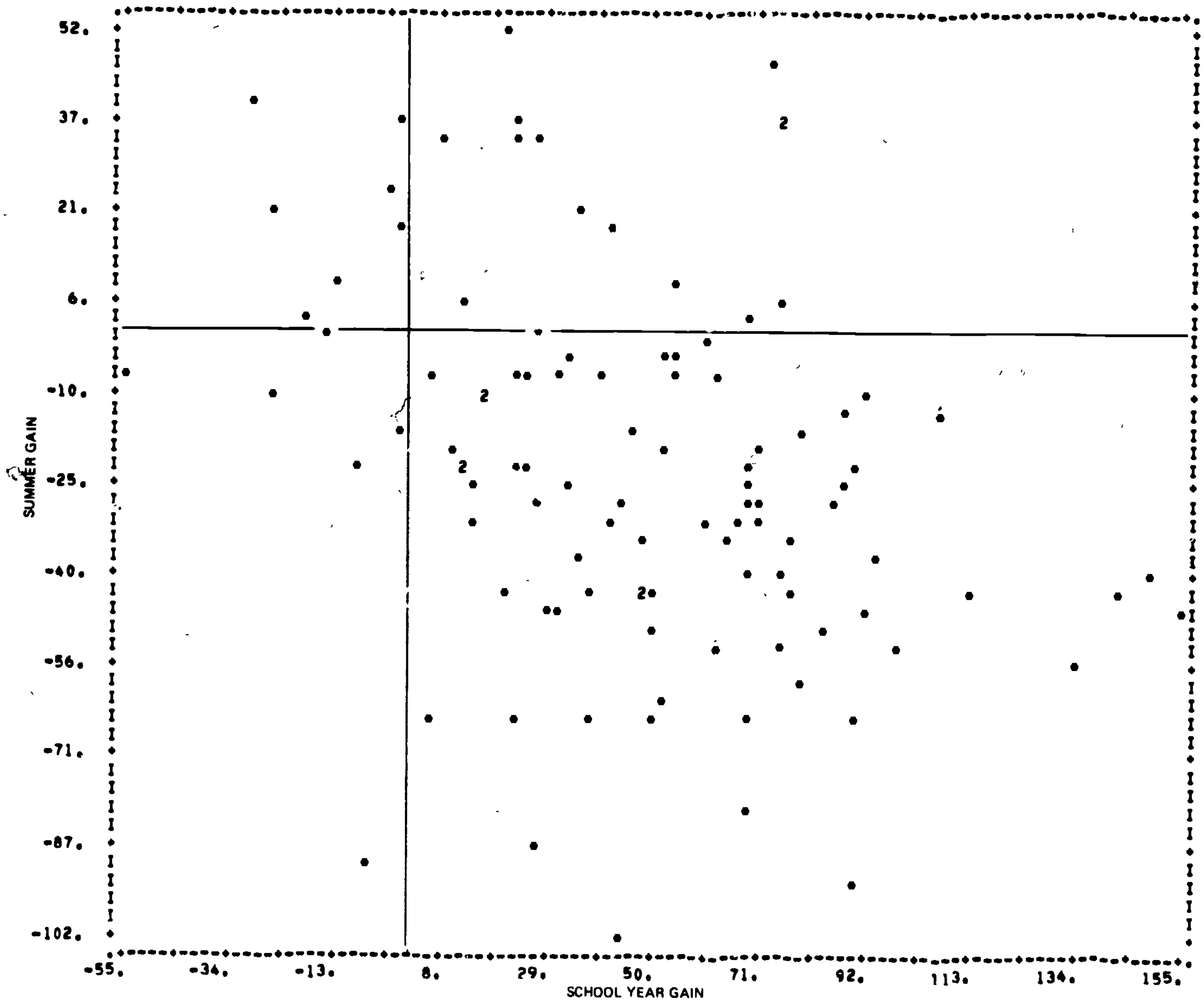

FIGURE V 1 RELATIONSHIP BETWEEN SCHOOL YEAR AND SUMMER GAINS IN CTBS STANDARD SCORE UNITS FOR LONG BEACH DPM STUDENTS GRADE 7, COHORT 3, SCHOOL A

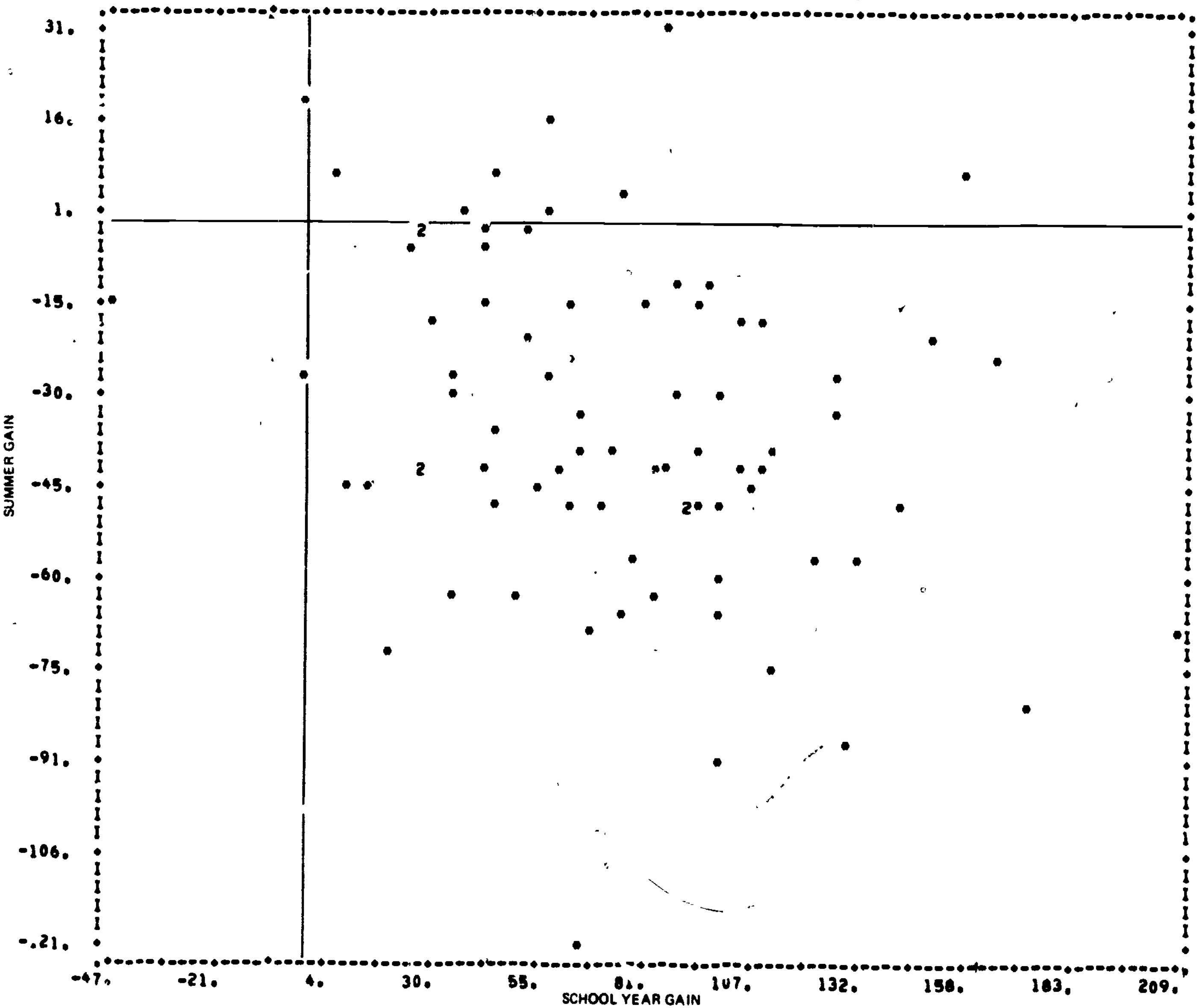

FIGURE V-2 RELATIONSHIP BETWEEN SCHOOL-YEAR AND SUMMER GAINS IN CTBS STANDARD SCORE UNITS FOR LONG BEACH DPM STUDENTS GRADE 7, COHORT 4, SCHOOL A

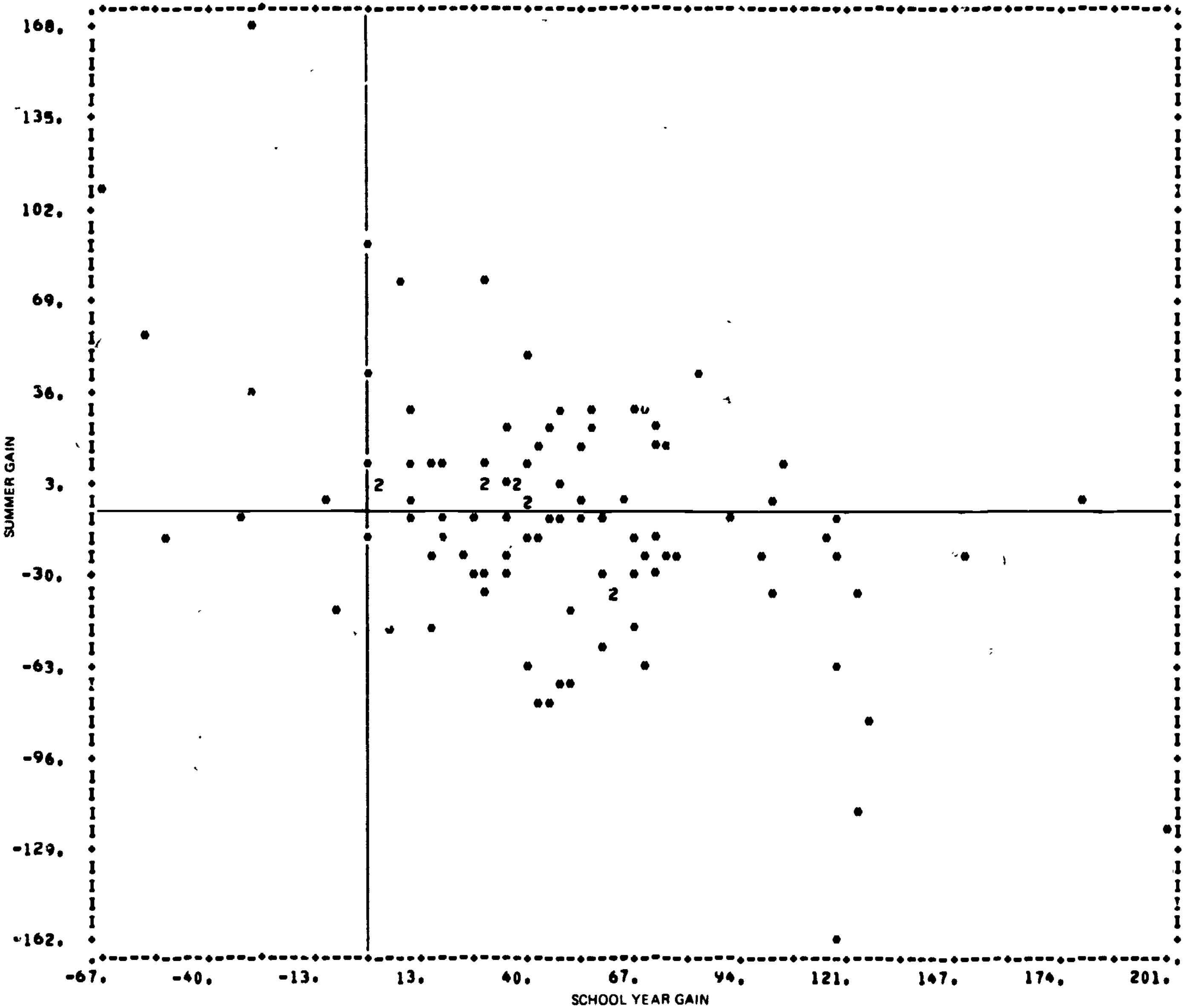

FIGURE V 3 RELATIONSHIP BETWEEN SCHOOL YEAR AND SUMMER GAINS IN CTBS STANDARD SCORE UNITS FOR SANTA BARBARA DPR STUDENTS GRADE 7

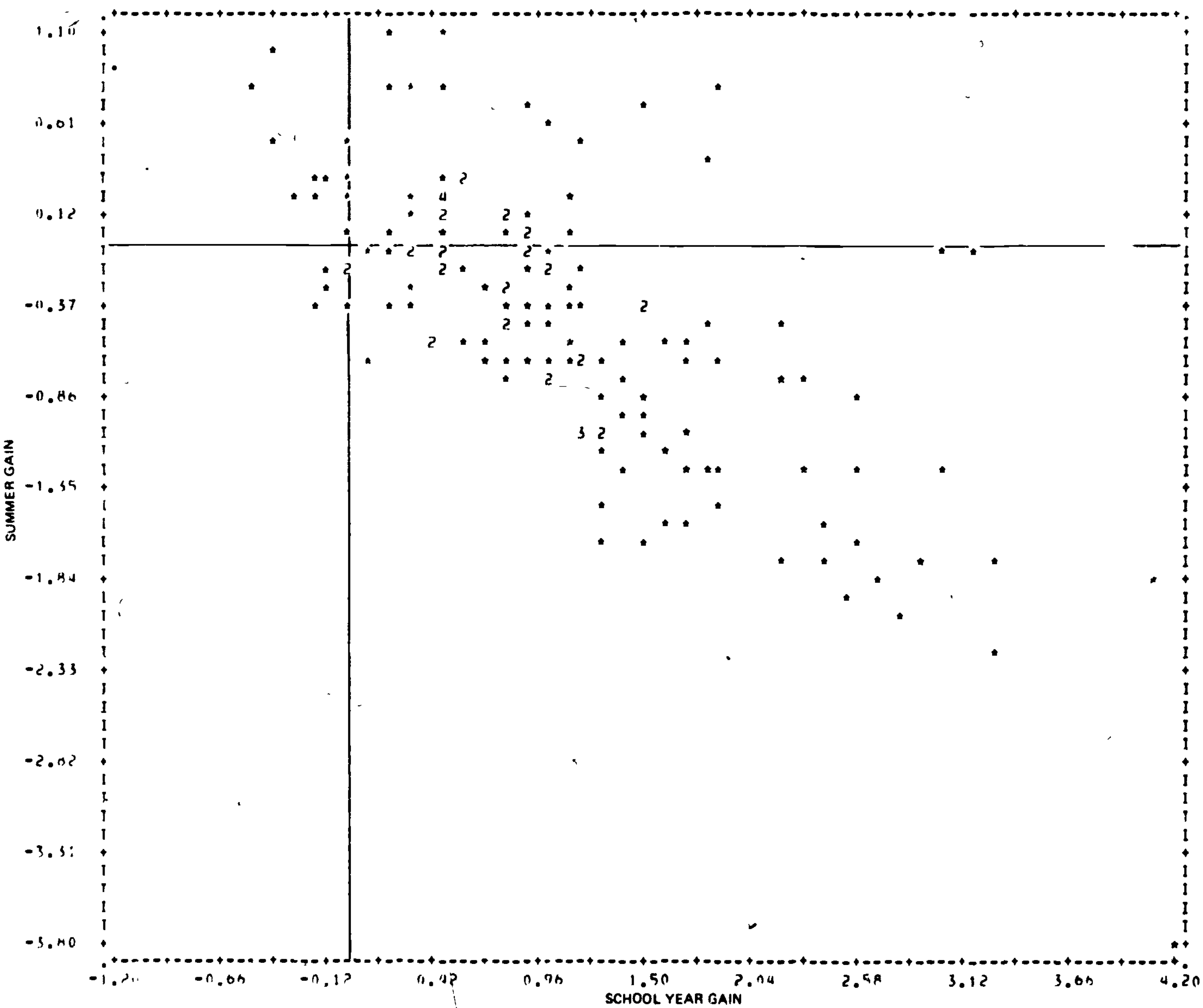

FIGURE V-4 RELATIONSHIP BETWEEN SCHOOL-YEAR AND SUMMER GAINS IN GATES-MacGINITIE GRADE EQUIVALENTS FOR OMAHA HILINC STUDENTS GRADE 3, COHORT 1

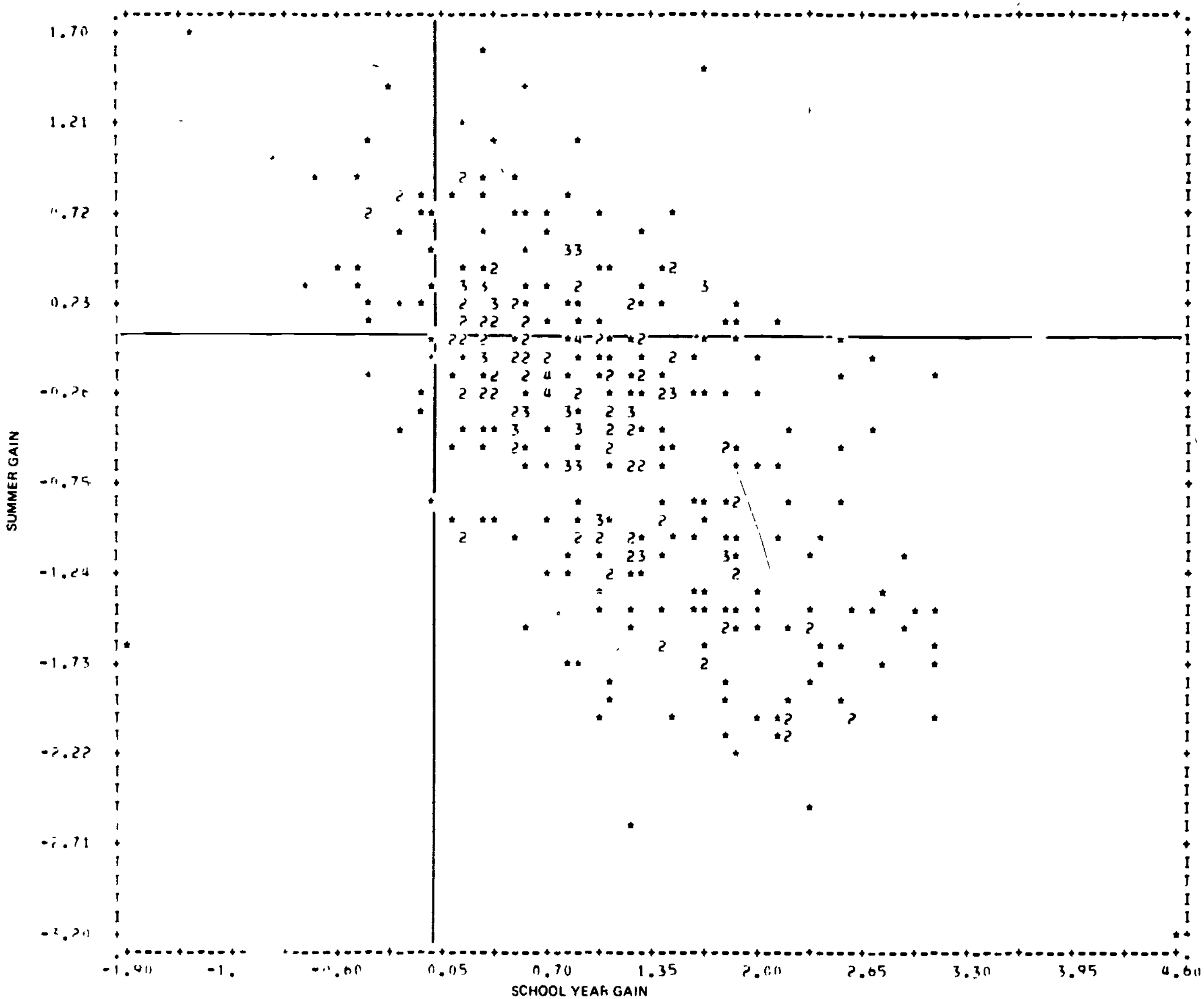

FIGURE V 5 RELATIONSHIP BETWEEN SCHOOL-YEAR AND SUMMER GAINS IN GATES-MacGINITIE GRADE EQUIVALENTS FOR OMAHA HILINC STUDENTS GRADE 4, COHORT 2

whose scores fall in the lower right-hand quadrant are those with gains over the school year and losses over the summer. Table V-2 following the scatterplots summarizes these numbers for each sample.

Table V-2

APPROXIMATE NUMBER OF STUDENTS WHOSE PATTERN FOLLOWS THE MEAN

Sample	Summer Loss		Summer Loss and School Year Gain	
	Number	Percent	Number	Percent
Long Beach DPM, School A				
Cohort 3 (n=109)	85	78%	80	73%
Cohort 4 (n=82)	73	89	71	87
Santa Barbara DPR (n=102)	56	55	52	51
Omaha HILINC				
Cohort 1 Grade 3 (n=152)	108	71	105	69
Cohort 2 Grade 4 (n=387)	258	67	251	65

The findings are encouraging in terms of generalizing findings at the mean level to individuals. In all five samples, at least 50% of the students follow the pattern of the mean. In four of the five samples the proportion of students with school-year gains and summer losses is at least 65%. The sample with the lowest percentage following the pattern, Santa Barbara with only 51%, is also the sample with the smallest summer loss at the mean level (see Table V-1). Therefore, we conclude that the phenomenon of summer losses at the mean level is not the result of a small number of extreme cases but rather reflects the pattern of the majority of students in each sample.

School-Year Gain Versus Summer Gain

We were next interested in whether any relationship existed between the amount of gain achieved during the school year and the amount lost over the summer. In other words, were students with large school-year

gains more or less likely than those with small school-year gains to have large losses over the summer. The scatterplots already presented in Figures V-1 through V-5 suggest that there is such a relationship; students who gain a lot over the school year tend to lose a lot over the summer and, conversely, those who gain little over the school-year lose little over the summer. These relationships are summarized by the correlation coefficients in Table V-3.

Table V-3

CORRELATION BETWEEN SCHOOL-YEAR GAINS AND SUMMER GAINS

Sample	Correlation Coefficient
Long Beach DPM	
Cohort 3 (n=109)	-.34
Cohort 4 (n=82)	-.30
Santa Barbara DPR (n=102)	-.49
Omaha HILINC	
Cohort 1 (n=152)	-.72
Cohort 2 (n=387)	-.60

All the correlations are negative (and statistically significant at the .01 level), indicating that large school-year gains tend to be associated with large summer losses and, the converse. Since these correlations are between two nonindependent gain scores (Spring minus Fall 1 and Fall 2 minus Spring), they are necessarily fraught with error. However, the size of the correlations suggests that there might be a real relationship, albeit inflated by measurement error. To determine if this were the case, we calculated a rough estimate of the expected correlation between two gains based on error alone. That is, we assumed that there was no correlation between school-year and summer gains and calculated the correlation using an estimate of the reliability of the tests. These calculations resulted in a correlation of approximately -.2 on the assumption of no true relationship.

Consequently, we concluded that our correlations, all of which exceed (in absolute sense) -.2, do reflect a true relationship, but probably one that ranges from -.1 to -.5 instead of -.3 to -.7. Nevertheless, in the world of educational research, correlations as high as -.5 are rare.

School-Year Gain Versus 12-Month Gain

If students with large school-year gains are likely to have large summer losses while those with small school-year gains will have small summer losses, then the next question of interest is whether the differences in amount of summer loss are substantial enough to alter a student's relative position by the end of the summer. That is, is the percentage of school-year gain that is lost higher for students with large school-year gains than for those with low school-year gains?

If this were true, judging students on the basis of a spring score would be tremendously misleading--not only because the fall score would be lower but also because the relative ranking of students would change. If this is not the case, however, the ranking of students would remain the same--those with the highest school-year gains would also have the highest 12-month, fall-to-fall gains. The high school-year gainers might lose more than low school-year gainers, but have more to lose; thus, they might remain at the top of the distribution of 12-month, fall-to-fall gains.

To test this hypothesis, we performed two similar analyses. First, we divided the school-year gains into seven intervals for the same five samples analyzed in the preceding discussion. We then calculated the mean of the 12-month (fall-to-fall) gains for all students falling in each interval. The results for Cohorts 3 and 4 in Long Beach, School A and for Santa Barbara are presented in standard scores in Table V-4, along with the grade-equivalent results for Cohort 1, Grade 3 and Cohort 2, Grade 4 in Omaha.

For all five samples there is a clear trend for the fall-to-fall gains to increase as the school-year gains increase. With a few minor

Table V-4

MEAN GAIN OVER 12 MONTHS (FALL TO FALL) BY SIZE OF SCHOOL-YEAR GAIN

Sample	Mean 12-Month Gain in Standard Scores by School Year Gain						
	≤ 0	1-20	21-40	41-60	61-80	81-100	≥101
Long Beach DPM, School A							
Cohort 3	-13.3 (n=13)	1.0 (n=14)	11.6 (n=22)	19.8 (n=18)	49.2 (n=25)	51.7 (n=11)	94 (n=6)
Cohort 4	-22.0 (n=3)	-25.2 (n=4)	6.9 (n=12)	29.9 (n=15)	[illegible]7.6 (n=13)	5[illegible].1 (n=17)	91.7 (n=18)
Santa Barbara DPR	14.8 (n=12)	14.4 (n=14)	28.9 (n=19)	38.9 (n=23)	58.5 (n=17)	96.3 (n=3)	81.7 (n=14)

	Means Gain in Grade-Equivalents						
	≤ 0	0.1-0.5	0.6-1.0	1.1-1.5	1.6-2.0	2.1-2.5	≥2.6
Omaha HILINC							
Cohort 1, Grade 3	-0.01 (n=18)	0.44 (n=30)	0.59 (n=36)	0.58 (n=32)	0.92 (n=15)	1.04 (n=8)	1.54 (n=13)
Cohort 2, Grade 4	-0.01 (n=32)	0.33 (n=75)	0.52 (n=98)	0.71 (n=81)	0.94 (n=56)	0.94 (n=28)	1.52 (n=17)

exceptions, these figures suggest a strong relationship between amount of school-year gain and amount of 12-month gain.

We then produced scatterplots of the relationship between school-year gain and 12-month gain to verify the findings from the first analysis. The scatterplots confirmed the relationship. They are summarized by the correlation coefficients shown in Table V-5, all of which are significant at the .001 level.

We conclude that even though students with large gains over the school year have large losses over the summer, the losses are not proportionately larger than those for students who have small gains over the school-year. Therefore, the ranking of students by size of gain at the end of the school year is similar to their ranking at the end of the following summer.

In conclusion, we find that the pattern of the means found in the analyses in Section IV is reflected at the individual level. This makes us feel more secure in making recommendations for evaluations conducted at a group level. Additionally, we suspect that there is an interesting relationship between amount of school year gain and amount of summer loss, as well as between amount of school-year gain and 12-month (fall-to-fall) gain. While our analyses are only a beginning in this line of investigation,* we think the initial findings are of sufficient interest to suggest pursuing this line of research.

---

*We also examined correlations between initial fall score and school-year gain but found no significant relationships.

Table V-5

CORRELATION BETWEEN SCHOOL-YEAR GAINS AND 12-MONTH (FALL TO FALL) GAINS

	Correlation Coefficient
Long Beach School A	
Cohort 3 (n=109)	0.69
Cohort 4 (n=82)	0.81
Santa Barbara (n=102)	0.52
Omaha HILINC	
Cohort 1 (n=152)	0.58
Cohort 2 (n=387)	0.48

VI SUMMARY AND RECOMMENDATIONS

Summary

Increasing the achievement of educationally disadvantaged students is a widely shared goal of compensatory education. This goal implies that increases in achievement can in some way affect the futures of disadvantaged children by equipping them with skills equivalent to those of their more advantaged peers. If increases in achievement are ephemeral, this goal will not be realized. Therefore, we have argued that judgments of the effectiveness of compensatory education programs should include measurement of the extent to which the program impact is lasting.

Only a very few studies of compensatory education have investigated the issue of sustained effects, and most of these are restricted to preschool programs. Since we could not draw on previous research, we turned our efforts to reanalyzing previously collected evaluation data; data that would allow estimates to be made of a sustained program impact. We obtained and analyzed data from four different compensatory education programs.

The primary finding of these analyses is that conclusions about program effectiveness, regardless of what standard is used, are greatly influenced by the period of time over which the program is judged. Specifically, we show that the inclusion of the summer months in the evaluation can substantially reduce estimates of achievement and often reverse positive judgments of program effectiveness. This results from the fact that losses in achievement often occur over the summer. In three of the four data sets presented, gains during the school year were followed by losses over the summer. In the fourth, although there was not an actual achievement loss over the summer, there was a reduction in rate of achievement.

Additionally, we demonstrate that different standards for success can result in different conclusions about program effectiveness. We have not explicitly compared the standards to each other since our primary interest was the effect of the time period for each standard. Nevertheless, we showed that the 10-percentile-point standard is more stringent than the two standards which entail grade-equivalent scores and thus is less likely to be met, especially during a 12-month, fall-to-fall evaluation.

Finally, the extent to which individuals in each sample follow the pattern discovered in the means was investigated. In the five samples studied, the achievement patterns of a majority of the individual students were the same as the pattern of the means. We conclude, therefore, that the consistent finding of school-year gains and summer losses is not a function of a small number of individuals in the sample with large summer losses.

As a last step, we looked at the relationship between the size of the school-year gain and the size of the summer gain (usually loss) for individuals. Although the correlations describing this relationship are fraught with measurement error, they were sufficiently large to convince us that there is an association between amount of school-year gain and summer loss. Specifically, students who gain the most over the school-year tend to be those who lose the most over the summer. However, analyses of the relationship between school-year gain and 12-month gain suggest that the ranking of students by size of gain does not shift dramatically from the end of one school year to the beginning of the next.

These data represent the only attempt to address the issue of summer loss with several longitudinal data sets, thereby eliminating the confounding introduced by cross-sectional data. Although we analyzed only four data sets, they represent different programs, different age levels, different subjects, different tests and many different schools. Since the findings of summer losses are quite consistent across all of these variables, we suspect that our conclusions are not limited to these four programs. Combined with questions raised by previous research, such

as the inconsistencies between school-year evaluation results and the results of annual state-wide testing programs, we suspect that the existence of summer losses is quite common for educationally disadvantaged students. Therefore, we urge that this phenomenon be taken into account in designing and carrying out evaluations of compensatory education programs.

It should be noted that our data demonstrate that programs can show evidence of sustained effects. Hence, a longer evaluation time period does not imply that all programs would be judged ineffec e.

Recommendations

ESEA Title I programs are usually evaluated on the basis of fall and spring test scores for a given year or a spring only scor (sometimes us g the previous spring score as a pretest). For districts that adm ster tests both fall and spring, our recommendation is not to change data collection practices, but rather to include analyses of students over the 12-month, fall-to-fall period. For districts that administer tests in the spring only, we suggest a change in data collection. If only one test can be administered annually, we recommend that this be done in the fall, thus permitting analyses of fall-to-fall achievement. When a program is evaluated on the basis of spring-to-spring scores, the results are perhaps not as misleading as those based on a fall-to-spring period since one summer is included in the spring-to-spring period of time. However, from a logical perspective, one should look for sustained gains some time after participation in the program. Therefore, evaluating a program from one spring to the next does not reflect the extent to which gains have been sustained after the program.

We are particularly concerned over the practice of "graduating" students from a program on the basis of a spring test score. When a district uses a spring score for determining program eligibility,*

---

*This practice exists in many districts, but we have no information on how widespread the practice is on a national scale.

students who attain their "expected" grade-level score are no longer eligible for program participation. Since some of the achievement gain reflected in the spring score may be lost by the end of the summer, extreme care should be taken in assuming that a spring score accurately reflects a student's achievement level. We urge, therefore, that fall scores rather than spring scores be used as a basis for judging eligibility for the program.

While we are willing to make suggestions concerning appropriate evaluation strategies, we are not willing to draw conclusions about the causes and therefore possible solutions for the summer loss phenomenon. Our recommendations are concerned with providing valuable information to program personnel about sustained achievement gains. We hope that this would be a first step in understanding why summer losses occur. If, for example, the phenomenon is a function of the measures used (the standardized achievement tests), one would want to change the measures instead of the program. If it is a result of instructional techniques that militate against retention, then the techniques should be changed. Without additional information on the causes, it is dangerous to suggest alternatives such as a different school calendar or summer school program. Therefore, the next step in this line of research should concentrate on explaining summer losses and relationships at the individual level. Only at this point can one recommend an appropriate remedy without the risk of exacerbating the situation.

REFERENCES

Baker, P., and S. H. Pelavin, 1976. "Issues of Reliability and Directional Bias in Standardized Achievement Tests: The Case of MAT 70," The Rand Paper Series, P-5689, Rand Corporation, Santa Monica, California.

Bronfenbrenner, U., 1974. "A Report of Longitudinal Evaluation of Pre-School Programs," U.S. Department of Health, Education and Welfare, Office of Human Development.

Bryk, A. S., and H. I. Weisberg, 1976. "Value-Added Analysis: A Dynamic Approach to the Estimation of Treating Effects," Journal of Educational Statistics, Vol. 1, pp. 127-157

Cronbach, L. J., and L. Furby, 1970. "How We Should Measure Change--or Should We?" Psychology Bulletin, Vol. 74, pp. 68-80.

Foat, C., 1974. "Selecting Exemplary Compensatory Education Projects for Dissemination Via Project Information Packages," Report No. UR-242, RMC Research Corporation, Los Altos, California.

Goodson, B.D. and R.D. Hess, 1976. "Parents as Teachers of Young Children: An Evaluative Review of Some Contemporary Concepts and Programs," Stanford School of Education, Stanford University, Stanford, California.

Goulet, Larry. et al., 1975. "Investigation of Methodological Problems in Educational Research: Longitudinal Methodology," Final Report, Contract No. NIE-C-74-0124, Project No. 4-1114, National Institute of Education, Washington, D.C.

Hayes, D. P. and J. Grether, 1969. "The School Year Vacations: When Do Students Learn?" Revision of a paper presented at the Eastern Sociological Association Convention, New York, New York.

Harris, C. W. (Ed.), 1963. Problems in Measuring Change. (University of Wisconsin Press, Madison, Wisconsin).

Heyns, B., 1977. "Exposure and the Effects of Schooling" Office of Economic Opportunity Grant No. 10301 (transferred to the National Institute of Education), in progress.

Murnane, R. J., 1975. The Impact of School Resources on the Learning of Inner City Children, Bollinger Publishing Company, Cambridge, Mass.

Pelavin, S. H. and P. Barker, 1976. "The Generalizability of the Results of a Standardized Achievement Test." The Rand Paper Series, P-5678, Rand Corporation, Santa Monica, California.

Pelavin, S. H., Jane L. David and Ann W. Porteus, 1976. "Synthesis of Research on the Effectiveness of Compensatory Education Programs: Interim Report for the National Institute of Education, Stanford Research Institute, Menlo Park, California.

Stearns, M. S., 1977. "Evaluation of the Field Test of Project Information Packages: Volume I Summary Report," Stanford Research Institute, Menlo Park, California.

Stearns, M. S., 1971. "Report on Preschool Programs: The Effects of Pre-School Programs on Disadvantaged Children and Their Families," U.S. Department of Health, Education and Welfare, Office of Child Development.

Tallmadge, G. K. and D. P. Horst, 1974. "A Procedural Guide for Validating Achievement Gains in Educational Projects." RMC research Corporation, Mountain View, California.

Thomas, T. C. and S. H. Pelavin, 1976. "Patterns in ESEA Title I Reading Achievement." Stanford Research Institute, Menlo Park, California.

White, S. H., et al., 1973. "Federal Programs for Young Children: Review and Recommendations." Volumes 1-3, The Huron Institute, Cambridge, Massachusetts.